文章编号：1003-0077(2017)03-0077-09

# 中文模糊限制信息范围语料库的研究与构建

周惠巍[1]，杨 欢[1]，徐俊利[1]，张 静[2]，亢世勇[2]

(1. 大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2. 鲁东大学 文学院，山东 烟台 264025)

**摘 要**：模糊限制语用于表示不确定性的观点。由模糊限制语所引导的信息为模糊限制信息，开展中文模糊限制信息检测研究，对事实信息抽取意义重大。模糊限制信息检测包含模糊限制性句子识别和模糊限制信息范围检测两个子任务。中文模糊限制信息范围语料库的缺乏，影响了中文模糊限制信息检测的研究。该文研究制定了基于短语结构的中文模糊限制信息范围标注规则，构建了中文模糊限制信息范围语料库。最后对标注的语料库进行了统计和分析。该文语料库的构建为中文模糊限制信息检测研究提供了资源支持。

**关键词**：中文模糊限制信息范围；标注规则；语料库

**中图分类号**：TP391 **文献标识码**：A

## Construction of Chinese Hedge Scope Corpus

ZHOU Huiwei[1], YANG Huan[1], XU Junli[1], ZHANG Jing[2], KANG Shiyong[2]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China;
2. School of Liberal Arts, Ludong University, Yantai, Shandong 264025, China)

**Abstract**: Hedge is usually used to express uncertainty. Hedge information indicates that authors do not backup their statements with facts. Chinese hedge information detection is of great significance for Chinese factual information extraction. Hedge information detection contains two subtasks: identifying hedges and detecting the in-sentence scopes of hedge cues. The lack of Chinese hedge scope corpus has limited the research of Chinese hedge scope information detection. This paper first manually crafted the syntactic rules for Chinese hedge scope annotation, and then constructs a Chinese hedge scope corpus. Finally, we statistically analyzed the corpus. The construction of the corpus provides a great support for Chinese uncertainty detection.

**Key words**: Chinese hedge scope; annotation rules; corpus

## 1 引言

模糊性是人类语言的一种属性，由于各种局限性，在语言交流和写作中，常常借助模糊限制语(hedges)表达不确定性的含义[1]。由模糊限制语所引导的信息为模糊限制信息(hedge information)。开展模糊限制信息检测研究，对事实信息抽取具有重要意义。英文模糊限制语研究开始较早，Prince等[2]从语用功能上将模糊限制语分为变动型和缓和型。近年来，模糊限制信息检测研究引起了国内外研究人员的广泛关注。2010 年计算自然语言学会议(Conference on Natural Language Learning, CoNLL)提出了模糊限制语识别及其范围(scope)检测共享任务(share task)[3]。其中模糊限制语识别包含生物医学和维基百科两个领域。生物医学领域语料源自 BioScope 语料库[4]，维基百科语料源自 WikiWeasel 语料库[3]，各两万句。模糊限制信息范围检测只采用了生物医学领域的 BioScope 语料库[4]。该语料库按模糊限制语的词性制定了范围标注规则。公开发表的英文模糊限制信息语料库还有新闻领域的 FactBank 语料库[5]。模糊限制信息语料库的构建促进了英文模糊限制信息检测的研究[6]。

近年来，模糊限制信息检测研究引起了国内研

收稿日期：2015-09-28 定稿日期：2016-02-03
基金项目：国家自然科学基金(61272375)

究人员的广泛关注[7-8]。邹博伟等[7]详细阐述了CoNLL-2010共享任务及不确定信息研究现状,并指出语料库的构建是中文模糊限制信息研究的重要基础。周惠巍等[8]基于句法结构约束检测模糊限制信息范围,在CoNLL-2010共享任务数据集上取得了较好的检测性能。

中文模糊限制语也广泛地用于生物医学等各个领域[9-10]。如例(1)源自生物医学领域文献,作者使用模糊限制语"可能",表明命题"这是由于增加了AKT,ERK磷酸化而引起的"的不确定性。而例(1)前半部分"在C6细胞中,血清的存在正调控内源受体的激活,使细胞凋亡率降低"为事实信息。因此在检测模糊限制信息中,模糊限制语识别及其范围检测同样重要。

例1 在C6细胞中,血清的存在正调控内源受体的激活,使细胞凋亡率降低,(这可能是由于增加了AKT,ERK磷酸化而引起的)$_{scope}$。

与英文相比,中文模糊限制信息检测研究开始较晚。何自然[9]在Prince等人[2]的研究基础上,将变动型模糊限制语分为程度变动型和范围变动型,将缓和型模糊限制语分为直接缓和型和间接缓和型,但是没有研究语料库的构建。Chen等人[11]构建了一个中文模糊限制语及其范围语料库,包含《计算机学报》论文4 842句,然而文中仅指出了副词和动词性模糊限制语的限制范围标注规则。曹媛等[12]在ACE2005中文事件抽取语料库上,根据事件选择谓词的语义,标注了事件的事实性程度,包括"确定"、"可能"和"不确定"三种取值。该语料可以用于事件的事实性研究。计峰等人[13]在新闻领域标注了一万句语料,进行中文不确定句子识别研究,该语料仅标注了模糊限制语,没有标注其限制范围。Zou等人[14]在科技文献、股市和产品评论三个领域,构建了16 841句模糊限制语及其范围语料,根据上下文语义,标注模糊限制语,基于完整性和连续性原则标注模糊限制信息范围,没有阐述具体的标注规则。

中文医学文献包含大量模糊限制语[15]。除医学文献外,维基百科作为一个用户协作编辑的知识系统,其中蕴涵了丰富的信息,成为信息抽取的重要语料资源。但是当撰写者不能提供完全准确的信息时,往往使用模糊限制语,使自己的陈述更客观。本文在生物医学和维基百科两个领域,根据模糊限制语的类型、词性及句子的短语结构,制定了详细的中文模糊限制信息范围标注规则,并构建了模糊限制信息范围语料库。

本文组织结构如下:第二节阐述了中文模糊限制语的分类和及其范围标注规则,并描述了标注的过程;第三节对标注完成的语料进行了统计和分析;第四节是结论与展望。

## 2 中文模糊限制信息范围语料库的设计与构建

### 2.1 中文模糊限制语分类

根据Prince等人[2]和何自然[9]的分类方法,模糊限制语可分为变动型和缓和型两类。在此基础上,本文根据模糊限制语的语义和语用功能,对这两大类模糊限制语进行了更细致的划分。

(1) 变动型模糊限制语。

变动型模糊限制语是对话题本身进行某种程度的限制,它能修改话题原来的真值。根据话题变动的类型,此类模糊限制语可细分为数量变动、程度变动、范围变动和频率变动四个类型。

• 数量变动型:当说话人不能明确地说出具体的数字,但是能估计出一个大概的数量时,使用到数量变动模糊限制语。如:"少数"、"大部分"等。

• 程度变动型:将一些接近正确但不敢肯定完全正确的话题,表述得与实际情况更接近,避免过于武断,表明话题与真实情况的接近程度。如:"有点"、"稍微"等。

• 范围变动型:可以在一定的范围内理解话题的意义,而不必考虑具体情况与所说的话题的接近程度。如:"大约"、"在一定范围内""将近"等。

• 频率变动型:用于反映一个事件发生的频率。如:"常常"、"偶尔"等。

(2) 缓和型模糊限制语。

当说话人提出某一个论断时,缓和型模糊限制语可以缓和说话人的语气,减轻说话人为此论断所负的责任,这类模糊限制语不改变话题的真值。根据缓和型模糊限制语的语用功能将其细分为主观见解型、客观依据型、探知结论型和条件假设型四类。

• 主观见解型:用来表示说话人阐述的话题只是个人的主观见解。使用这类模糊限制语可以在一定程度上削弱说话人对话题所承担的责任。如:"我认为"、"就我所知"等。

• 客观依据型:借助第三方或大家普遍认同的观点,表达说话人对某事所持有的态度。说话人

在一定程度上同意第三方的观点，只是他对此观点究竟有多大程度的赞同，在话语中看不出来，只能另作推断。例如，“据说”、“有人说”等。

• 探知结论型：用来表示对某个结论的推测，根据存在的现象推知未来可能会发生的事情或有待证明的结论。如：“表明”、“可能”和“仍不清楚”等。

• 条件假设型：通过给出假定的前提条件表明说话人的意愿，但现在事实是怎样的并不知晓。如：“如果”、“假定”等。

### 2.2　标注过程

为了保证标注的准确性，首先由规则制定者给两名独立标注人员讲解标注规则，并共同讨论、修正规则。然后由两名独立标注人员根据规则，分别标注模糊限制语及其范围。最后规则制定者统一两份标注语料中不一致的标注，形成最终的中文模糊限制信息范围语料库。具体的标注过程如图 1 所示。

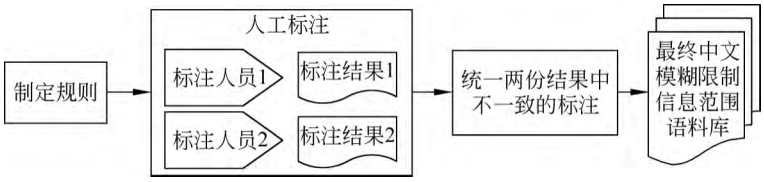


图 1　语料标注过程

### 2.3　基本标注规则

模糊限制语的标注遵循“最小原则”：标注能表明模糊限制性的最小单元为模糊限制语，多个模糊限制语组合起来表示模糊限制性时，分别标注每个模糊限制语。中文模糊限制范围标注遵循以下基本原则：

（1）连续性和完整性原则。

中文模糊限制信息范围标注遵循“连续原则”，即模糊限制语的作用范围为包含该模糊限制语的一段连续字符串。同时保持“完整性”，即为包含该模糊限制语的具有完整语义的最大句法单元。模糊限制信息虽具有不确定性，但也是有价值的信息，可用于知识发现等[16]。所以应该尽量完整地标记出来。这与英文 BioScope 语料库[4]的标注原则不同，BioScope 标注了每个模糊限制语的语法修饰范围。

（2）当模糊限制语为动词的被动语态时，模糊限制信息范围应该包含主语。

例句(2)中的模糊限制语“被认为”是动词的被动语态，如果其模糊限制信息范围仅为其修饰的“被认为能透露出受测者是否说谎”，不包含动词的主语“说谎而引起此类生理反应的变化”，则无法表示完整语义。所以应该包含动词的主语，这也符合“完整性”。

例 2　由于此类生理反应是不由自主地产生的，＜scope＞说谎而引起此类生理反应的变化＜ccue＞被认为＜/ccue＞能透露出受测者是否说谎＜/scope＞。

我们还规定，如果模糊限制信息范围结束于句尾，则不包含句尾的标点符号。如果一个句子中有多个模糊限制语，各个模糊限制信息范围可以并列，也可以嵌套，但不能存在交叉。

### 2.4　具体标注规则

不论是中文还是英文，模糊限制信息范围的界定都具有依赖于句法结构的特点[17-18]。根据模糊限制语的类型、词性及句子短语结构，制定模糊限制信息范围标注规则。模糊限制语可以分为变动型和缓和型两大类。缓和型模糊限制语中，探知结论型与客观依据型居多，并且，探知结论型模糊限制语大多是动词和副词。变动型的模糊限制语大多是形容词及副词。下面介绍这些常见类型的模糊限制语的范围标注规则，并采用斯坦福句法解析器（Stanford parser）①，获得例句的短语结构树，辅助对规则的理解。

（1）缓和型模糊限制语的限制范围标注规则。

• 动词性探知结论型模糊限制语：其范围为距离模糊限制语最近的祖先动词短语(verb phrase，VP)。

例 3　苏联的科学家 Bukasov(1935)和 Vavilov(1935)＜scope＞＜ccue＞推断＜/ccue＞欧洲的马铃薯的起源就是智利的马铃薯＜/scope＞。

例 3 中使用动词性探知结论型模糊限制语“推断”，说明命题“欧洲的马铃薯的起源就是智利的马铃薯”是一个不确定的、待证明的命题，而前面的“苏联的科学家 Bukasov(1935)和 Vavilov(1935)”是一个确定的信息。并且不加入这一确定的信息，“推

① http://nlp.stanford.edu/software/lex-parser.shtml

断”所引导的动词短语“推断欧洲的马铃薯的起源就是智利的马铃薯”即可表示完整的语义。这与例2的被动语态不同。

例3的短语结构树如图2所示。$VP_1$是距离模糊限制语“推断”最近的祖先VP类型节点。模糊限制信息范围为该$VP_1$结构，即$VP_1$的第一个词“推断”为范围左边界，$VP_1$的最后一个词“马铃薯”为范围右边界。

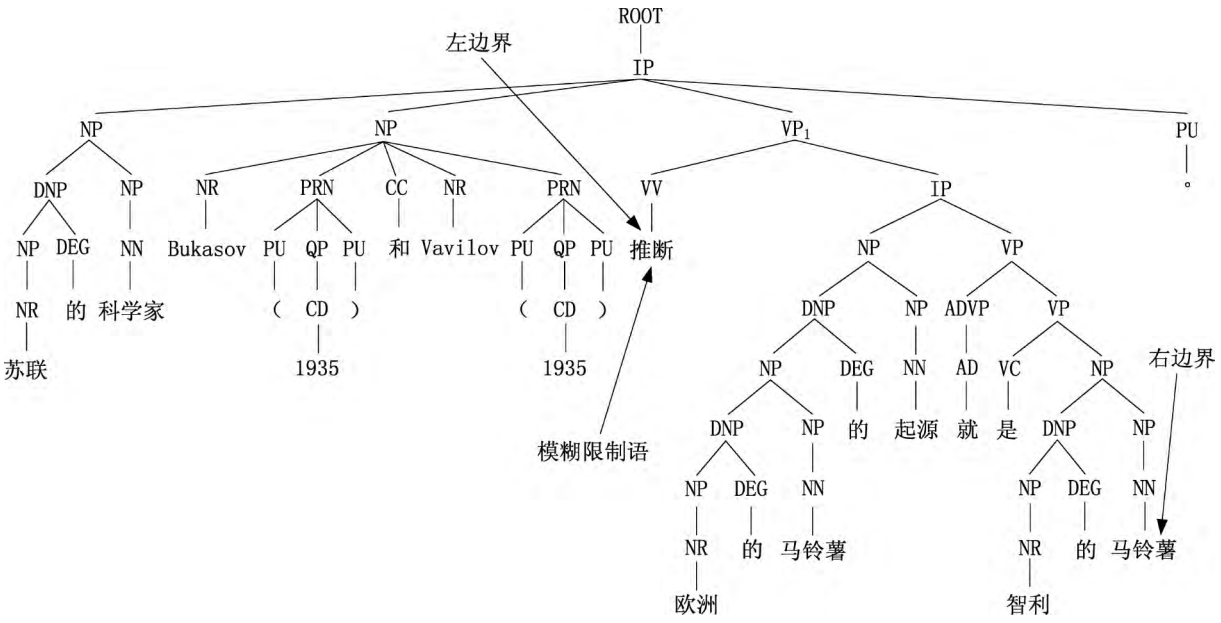


图2 例3的短语结构树

• 副词性探知结论型模糊限制语：其范围包括距离模糊限制语最近的祖先VP，以及与该VP类型节点同层次的左边相邻的名词短语(noun phrase，NP)。

例4 结论：＜scope＞TLR＜ccue＞可能＜/ccue＞通过抑制p38MAPK-FN通路对糖尿病肾病大鼠的肾脏产生保护作用＜/scope＞。

例4中，使用“可能”，使得命题“TLR通过抑制p38MAPK-FN通路对糖尿病肾病大鼠的肾脏产生保护作用”具有不确定性。因为“TLR”是“可能”的主语，为标注完整的语义信息，“TLR”应该包含在模糊限制信息范围内。

例4的短语结构树如图3所示。$VP_1$是距离“可能”最近的祖先VP类型节点，$NP_1$是与$VP_1$同层次的左边相邻的NP结构。所以$NP_1$的第一个词“TRL”为范围左边界，$VP_1$的最后一个词“作用”为范围右边界。

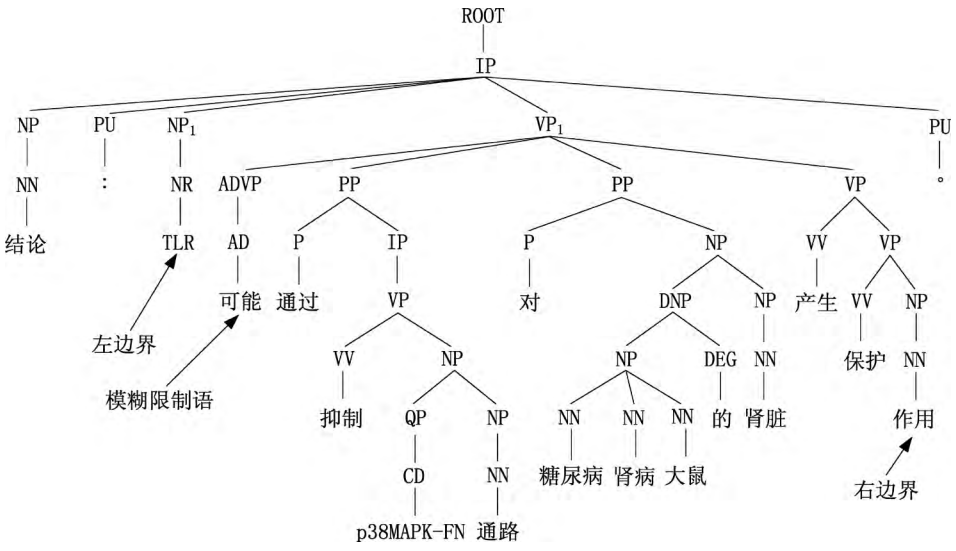


图3 例4的短语结构树

• 客观依据型模糊限制语：其范围包含离模糊限制语最接近的祖先介词短语（preposition phrase，PP），以及与该 PP 类型节点同层次的右侧最接近的 VP 结构。

例 5　＜scope＞＜ccue＞根据＜/ccue＞胎儿源性成人疾病学说，IUGR 会明显增加成人后患心血管疾病的概率＜/scope＞。

例 5 中，作者使用“根据”，减轻对命题“IUGR 会明显增加成人后患心血管疾病的概率”真假性所负的责任。所以该命题应该包含在模糊限制信息范围内。基于“连续性原则”、“根据”和命题之间的“胎儿源性成人疾病学说”也应该包含在模糊限制信息范围内。

（2）变动型模糊限制语的限制范围标注规则。

• 形容词性变动型模糊限制语：当模糊限制语为形容词时，它通常在一个 NP 结构中。如果距离模糊限制语最近的祖先 NP 类型节点的父亲节点是 VP 类型，则模糊限制信息范围为包含模糊限制语的连续最上层祖先 VP 结构，以及与该 VP 类型节点同层次的左边相邻的 NP 结构，如例 6。如果距离模糊限制语最近的祖先 NP 类型节点的父亲节点不是 VP 类型，则模糊限制信息范围包含该 NP 结构，以及与该 NP 类型节点同层次的右侧最接近的 VP 结构，如例 7。这与英文 BioScope 语料库[4]的标注原则不同，BioScope 认为形容词性模糊限制语的范围为其所修饰的名词短语。本文强调具有模糊性的完整命题。

例 6　总之，＜scope＞Toll 信号通路对中枢神经系统疾病有＜ccue＞一定的＜/ccue＞调控作用＜/scope＞。

例 6 的短语结构树如图 4 所示。其中，$NP_2$ 是距离“一定的”最近的祖先 NP 类型节点，$VP_1$ 是“一定的”连续的最上层祖先 VP 类型节点。连续的最上层祖先 VP 类型节点是指：如果离模糊限制语最近的祖先 VP 类型节点的父亲节点属性也是 VP，则继续沿着祖先节点路径向上寻找，直到找到父亲节点不是 VP 类型的最上层 VP 类型节点。此句中，距离模糊限制语“一定的”最近的祖先 VP 类型节点是 $VP_2$，沿着虚线向上寻找，找到 $VP_1$，沿着虚线继续向上寻找，发现 $VP_1$ 的父亲节点类型为 IP，返回到 $VP_1$。$NP_1$ 是与 $VP_1$ 同层次的左侧相邻的 NP 结构。所以 $NP_1$ 的第一个词“Toll”为范围左边界，$VP_1$ 的最后一个词“作用”为右边界。

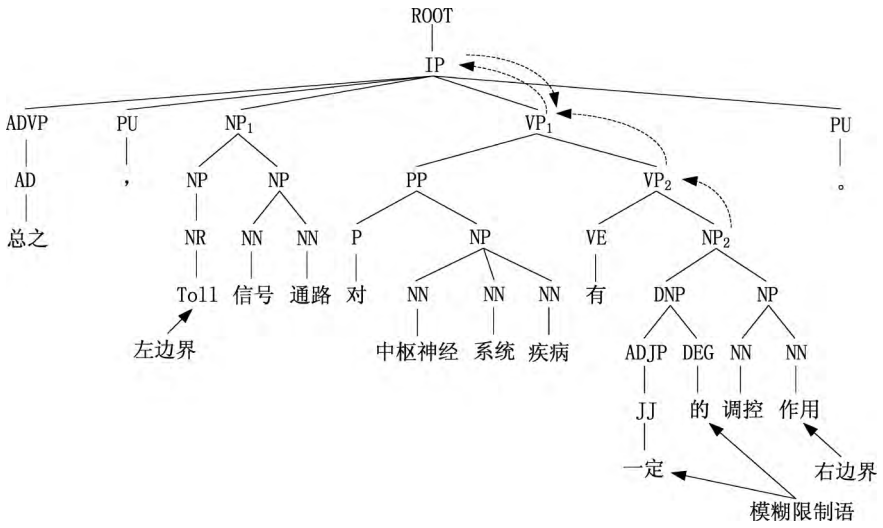


图 4　例 6 的短语结构树

例 7　修宪后＜scope＞国大的＜ccue＞大部分＜/ccue＞职责已经转交立法院＜/scope＞，其中也包括了弹劾总统的权力。

例 7 的短语结构树如图 5 所示。其中，$NP_1$ 是距离“大部分”最近的祖先 NP 类型节点，而 $NP_1$ 的父亲节点 IP 不是 VP 类型节点。所以模糊限制信息范围包含 $NP_1$ 结构，以及与该 $NP_1$ 类型节点同层次的右边最接近的 $VP_1$ 结构。

• 副词性变动型模糊限制语：模糊限制信息范围为包含模糊限制语的连续的最上层祖先 VP 结构，以及与该 VP 类型节点同层次的左边相邻的 NP 结构。

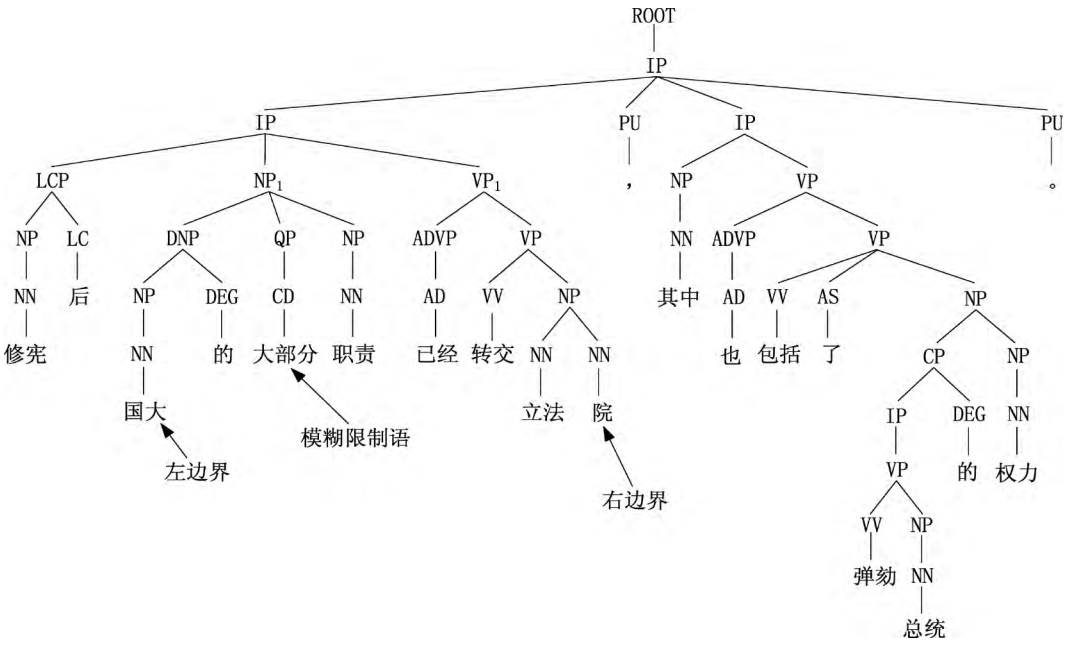


**图5 例7的短语结构树**

**例8** 1919年1月5日,<scope>红军进入明斯克,<ccue>几乎</ccue>没有遭遇抵抗</scope>,短命的白俄罗斯人民共和国垮台。

例8中,“几乎”使得“没有遭遇抵抗”的程度不确定,所以“没有遭遇抵抗”要包含在范围内,而“没有遭遇抵抗”的主语是“红军进入明斯克”这一动作,如果缺少该主语则命题不完整,所以“红军进入明斯克”也要包含在范围内。

中文使用千变万化,不是所有的句子都能基于规则进行标注。在实际标注过程中,需要根据模糊限制语的上下文和句子的语义标注模糊限制信息范围。另外,斯坦福句法解析错误较多,需要人工修正错误的句法解析结果。

## 3 数据统计及一致性分析

### 3.1 语料库的统计数据

在生物医学和维基百科两个领域共标注语料24 000余句。中文模糊限制语的统计信息如表1所示。生物医学文献中,33.30%的句子包含模糊限制信息。其中,48.03%的模糊限制语为变动型,51.97%为缓和型。维基百科中,33.10%的句子包含模糊限制信息。其中,71.99%的模糊限制语为变动型,28.01%为缓和型。可见两种类型的模糊限制语广泛地用于中文文献。而英文生物医学领域的BioScope语料库[4]仅标注了缓和型模糊限制语,WikiWeasel语料库[3]仅标注了变动型模糊限制语。

**表1 中文模糊限制语的统计信息**

| 统计条目 \ 文献领域 | 生物医学 | 维基百科 |
|---|---|---|
| 模糊限制性句子比例/% | 33.30 | 33.10 |
| 变动型模糊限制语比例/% | 48.03 | 71.99 |
| 缓和型模糊限制语比例/% | 51.97 | 28.01 |

中文模糊限制信息范围标注的统计信息如表2所示。从表2可以看出,模糊限制信息范围“不开始于模糊限制语”的数量多于“开始于模糊限制语”的数量。这主要是因为基于完整性,常常将主语也包含在范围内。“不结束于句尾”的数量多于“结束于句尾”的数量。然而,“开始于模糊限制语”和“结束于句尾”还是占有较大比例。

### 3.2 一致性分析

对每个模糊限制语都标记了唯一的范围开始和结束标记,所以召回率为百分之百。采用准确率作为一致率,分析标注一致性。中文模糊限制信息范围语料库的一致率如表3所示。Left-Scope为左边界匹配的一致率,Right-Scope为右边界匹配的一致率,Full-Scope为左、右边界同时匹配的一致率。各单元格中的第一项表示两份独立标注的语料间的一致率,第二项和第三项分别表示两份独立标注语料与最终语料间的一致率。

表 2　中文模糊限制信息范围的统计信息

| 统计条目 \ 文献领域 | 生物医学 | | | | | | 维基百科 | 生物医学和维基百科 |
|---|---|---|---|---|---|---|---|---|
| | 摘要 | 实验结果 | 讨论 | 结论 | 全文 | 生物医学总体 | | |
| 句子平均长度(＃词语) | 35.18 | 38.78 | 38.50 | 37.51 | 39.40 | 37.20 | 26.46 | 35.36 |
| 范围平均长度(＃词语) | 16.20 | 13.00 | 16.33 | 14.79 | 16.97 | 15.69 | 12.43 | 15.10 |
| ＃开始于模糊限制语 | 800 | 636 | 1 969 | 90 | 218 | 3 713 | 737 | 4 450 |
| ＃不开始于模糊限制语 | 1 321 | 873 | 2 252 | 250 | 227 | 4 923 | 1 161 | 6 084 |
| ＃结束于句尾 | 1 228 | 584 | 1 833 | 151 | 172 | 3 968 | 943 | 4 911 |
| ＃不结束于句尾 | 893 | 925 | 2 388 | 189 | 273 | 4 668 | 955 | 5 623 |

表 3　中文模糊限制信息范围语料标注的一致率

| | Left-Scope/% | Right-Scope/% | Full-Scope/% |
|---|---|---|---|
| 摘要 | 88.50 / 93.07 / 94.72 | 87.27 / 93.02 / 93.35 | 77.37 / 86.85 / 88.87 |
| 实验结果 | 78.53 / 84.03 / 91.32 | 90.26 / 92.05 / 94.83 | 71.37 / 79.39 / 88.73 |
| 讨论 | 93.51 / 95.62 / 96.73 | 91.87 / 94.08 / 96.68 | 86.33 / 90.64 / 94.08 |
| 结论 | 95.00 / 98.24 / 96.18 | 95.59 / 97.35 / 98.24 | 90.88 / 95.88 / 94.41 |
| 全文 | 95.28 / 96.63 / 98.20 | 95.73 / 97.53 / 97.30 | 91.46 / 94.83 / 95.73 |
| 维基百科 | 87.42 / 91.74 / 94.00 | 94.37 / 95.89 / 96.95 | 83.37 / 89.21 / 92.05 |

由表 3 可见，各单元格中的第一项均低于第二项和第三项，这是因为最终语料是规则的制定者对两份独立标注语料的不同之处进行统一后获得的，所以有可能和二者之一相同。Right-Scope 和 Left-Scope 的一致率十分接近，且 Right-Scope 的一致率略高于 Left-Scope，说明在标注过程中，界定中文模糊限制信息范围的左边界略难于中文模糊限制信息范围的右边界。Full-Scope 的一致率明显低于 Left-Scope 和 Right-Scope 的一致率。

表 4 是两份独立标注语料的 Full-Scope 一致率。本文对客观依据型和探知结论型模糊限制语制定了清楚的标注规则，从表 4 可以看出，每份语料中这两个类型的一致率都较高。可见制定准确的规则有助于中文模糊限制信息范围的标注。“实验结果”语料中的主观见解型模糊限制语的一致率为 0，这是因为该语料中只有两个主观见解型模糊限制语，而两名独立标注人员对这两个模糊限制语的范围标注都不一致。

表 4　两份独立标注语料的 Full-Scope 一致率

| | 程度变动/% | 范围变动/% | 频率变动/% | 数量变动/% | 主观见解型/% | 客观依据型/% | 探知结论型/% | 条件假设型/% |
|---|---|---|---|---|---|---|---|---|
| 摘要 | 80.27 | 70.55 | 82.50 | 74.64 | 60.00 | 75.00 | 77.52 | 77.78 |
| 实验结果 | 67.72 | 61.03 | 61.11 | 60.20 | 0.00 | 88.89 | 83.30 | 100.00 |
| 讨论 | 85.56 | 83.58 | 82.01 | 85.96 | 90.00 | 87.95 | 87.22 | 85.19 |
| 结论 | 97.37 | 87.50 | 85.71 | 92.59 | 83.33 | 100.00 | 83.96 | 83.33 |
| 全文 | 91.67 | 90.32 | 88.89 | 98.11 | 100.00 | 100.00 | 89.66 | 93.33 |
| 维基百科 | 83.29 | 78.95 | 84.49 | 84.27 | 87.50 | 86.97 | 80.87 | 66.67 |

虽然制定了清晰的标注规则，但仍存在标注分歧，说明标注存在一定的主观性，且中文语言丰富多彩，规则不能涵盖所有的情况。部分分歧如下：

（1）连接词是否要包含在模糊限制信息范围内。

标注（1）：陆地边界现在已清楚划定，并

＜scope＞＜ccue＞大略＜/ccue＞依据地理特征来界定＜/scope＞，例如：玻璃市河(PerlisRiver)、哥乐河(GolokRiver)与 PagalayanCanal。

标注(2)：陆地边界现在已清楚划定，＜scope＞并＜ccue＞大略＜/ccue＞依据地理特征来界定＜/scope＞，例如，玻璃市河(PerlisRiver)、哥乐河(GolokRiver)与 PagalayanCanal。

最终，我们按标注(1)进行统一，认为“并”和前面的句子有关系，对后面的句子没有影响，所以不将它包含在模糊限制信息范围内。

(2) 当一个句子中出现多个模糊限制语时，易出现标注分歧。

标注(1)：不会疼痛且没有感染的脸部肿胀也算是一种类型的腮腺炎，＜scope id＝" 1"＞它＜ccue id＝" 1"＞有可能＜/ccue id＝" 1"＞是＜scope id＝" 2"＞急性＜ccue id＝" 2"＞或＜/ccue id＝" 2"＞慢性的＜/scope＞＜/scope＞。

标注(2)：不会疼痛且没有感染的脸部肿胀也算是一种类型的腮腺炎，＜scope id＝" 1"＞＜scope id＝" 2"＞它＜ccue id＝" 1"＞有可能＜/ccue id＝" 1"＞是急性＜ccue id＝" 2"＞或＜/ccue id＝" 2"＞慢性的＜/scope＞＜/scope＞。

这个例句中有两个模糊限制语，第二个模糊限制语“或”的限制信息范围标注出现了不一致。这种情况下，为使两个模糊限制语的范围不重复，将标注(1)作为正确的标注。

### 3.3 与相关研究的比较

何自然[9]研究了模糊限制语的定义和分类，但没有进行语料库的构建研究。Chen 等人[11]构建了一个中文模糊限制语及其范围语料库，然而仅包含科学文献一个领域；指出了副词和动词的模糊限制范围应该扩展到从句或整个句子，但是没有阐明其他词性模糊限制语的范围标注规则。曹媛等人[12]在已有的中文事件抽取语料库上，根据谓词的语义，将事件划分为“确定”、“可能”和“不确定”三种。该语料可以用于事实性事件的抽取研究。计峰等人[13]为进行中文不确定句子识别研究，对 1 万句新闻领域语料进行了确定性和非确定性标注，但是没有标注模糊限制信息范围。Zou 等人[14]在科技文献、金融报道和产品评论三个领域，构建了模糊限制语及其范围语料；指出了模糊限制语及其范围标注的总原则，即根据上下文语义标注模糊限制语；基于完整性和连续性原则标注模糊限制信息范围。

本文根据模糊限制语的语义和语用功能，对模糊限制语进行了更细致的划分，使得模糊限制语的概念更加明确。且针对不同类型、不同词性的模糊限制语，详细阐述了其范围标注规则。详尽的标注规则，不但保证了标注语料的质量，对模糊限制信息范围检测研究，也具有指导意义。此外，本文在生物医学和维基百科两个领域，构建了模糊限制语及其范围语料，为模糊限制信息检测提供了充足的资源。

## 4 总结与展望

本文研究构建了生物医学和维基百科两个领域的中文模糊限制信息范围语料库。根据中文模糊限制语的类型、词性及句子的短语结构，制定了中文模糊限制信息范围标注规则。实验从语料的领域和模糊限制语的类别两个方面，统计了范围标注的一致性。基于详尽的标注规则和严格的标注过程，语料标注取得了较高的一致率。标注完成的语料库包含 10 534 个模糊限制语及其作用范围。语料规模足以用于中文模糊限制信息检测的研究。下一步我们将推出一个语料库的在线版本，为中文模糊限制语的研究提供共享资源。并根据使用者的反馈意见，继续完善标注规范，改进标注质量，扩大语料规模。

## 参考文献

[1] Lakoff G. Hedges: a study in meaning criteria and the logic of fuzzy concepts [J]. Journal of Philosophical Logic, 1973, 2(4): 458-508.

[2] Prince E F, Frader J, Bosk C. On hedging in physician-physician discourse [J]. Linguistics and the Professions, 1982: 83-97.

[3] Farkas R, Vincze V, Móra G, et al. The CoNLL 2010 Shared Task: Learning to detect hedges and their scope in natural language text [C]//Proceedings of the CoNLL, Uppsala, Sweden, 2010: 1-12.

[4] Vincze V, Szarvas G, Farkas R, et al. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes [J]. BMC Bioinformatics, 2008, 9(11): S9.

[5] Saurí R and Pustejovsky J. FactBank: A corpus annotated with event factuality [J]. Language Resources and Evaluation, 2009, 43(3): 227-268.

[6] Tang B Z, Wang X L, Wang X, et al. A cascade method for detecting hedges and their scope in natural language text [C]//Proceedings of the CoNLL, Uppsala, Sweden, 2010: 25-29.

[7] 邹博伟，周国栋，朱巧明. 否定与不确定信息抽取研究综述[J]. 中文信息学报，2015，29(4)：16-24.
[8] 周惠巍，杨欢，黄德根，等. 基于句法结构约束的模糊限制信息范围检测[J]. 中文信息学报，2013，27(5)：137-143.
[9] 何自然. 模糊限制语与言语交际[J]. 外国语(上海外国语学院学报)，1985，(5)：27-31.
[10] 贾晓凡，蒋跃. 基于小型语料库的模糊限制语分类方法的对比研究[J]. 外语艺术教育研究，2011，(3)：10-14.
[11] Chen Z C，Zou B W，Zhu Q M，et al. The scientific literature corpus for chinese negation and uncertainty identification [M]. Chinese Lexical Semantics. Springer Berlin Heidelberg，2013：657-667.
[12] 曹媛，朱巧明，李培峰. 中文事件事实性信息语料库的构建方法[J]. 中文信息学报，2013，27(6)：38-44.
[13] 计峰，邱锡鹏，黄萱菁. 中文不确定性句子的识别研究[C]. 全国信息检索学术会议，2010：594-601.
[14] Zou B W，Zhu Q M，Zhou G D. Negation and Speculation Identification in Chinese Language [C]//Proceedings of the ACL-2015，Beijing，2015：656-665.
[15] 陈萍，蒋跃. 中英医学论文摘要中模糊限制语的对比研究[J]. 外语艺术教育研究，2009，3(1)：15-20.
[16] Velldal E，Ovrelid L，Read J，et al. Speculation and negation：rules，rankers，and the role of syntax[J]. Association for Computational Linguistics，2012，38(2)：369-410.
[17] Cheng L X，Lin H F，Zhou F，et al. Enhancing the accuracy of knowledge discovery：a supervised learning method [J]. BMC Bioinformatics，2014，15(Suppl 12)：S9.
[18] Moncecchi G，Minel J，Wonsever D. The Influence of Syntactic Information on Hedge Scope Detection [C]//Proceedings of the 14th Ibero-American Conference on AI. Berlin：Springer，2014：83-94.

**周惠巍**(1969—)，博士，副教授，主要研究领域为生物医学信息挖掘、机器学习和自然语言处理。
E-mail：zhouhuiwei@dlut.edu.cn

**徐俊利**(1990—)，硕士研究生，主要研究领域为生物医学信息挖掘、机器学习和自然语言处理。
E-mail：xjlhello@ mail.dlut.edu.cn

**杨欢**(1988—)，硕士研究生，主要研究领域为生物医学信息挖掘、机器学习和自然语言处理。
E-mail：yanghuan_dlut@mail.dlut.edu.cn